Procedimiento de calculo de la materia probable de un documento de texto y su uso en el programa informatico *“Herramienta de Documentalista”*

Diago Marquez, Francisco Jose (Licenciado en Documentacion, Autor del Programa)

Resumen

Se muestra la forma de realizar los operaraciones y procesos para calcular la materia probable de un documento de texto a partir de su contenido. El calculo (un clasico en la indixacion y clasificacion automatica) se basa en la cantidad de palabras que aparecen y se repiten comparandose con un corpus de palabras asociados previamente a la materia que le corresponde. Los resultados por el momento son aceptables aunque es necesario refinar mas el procedimiento.

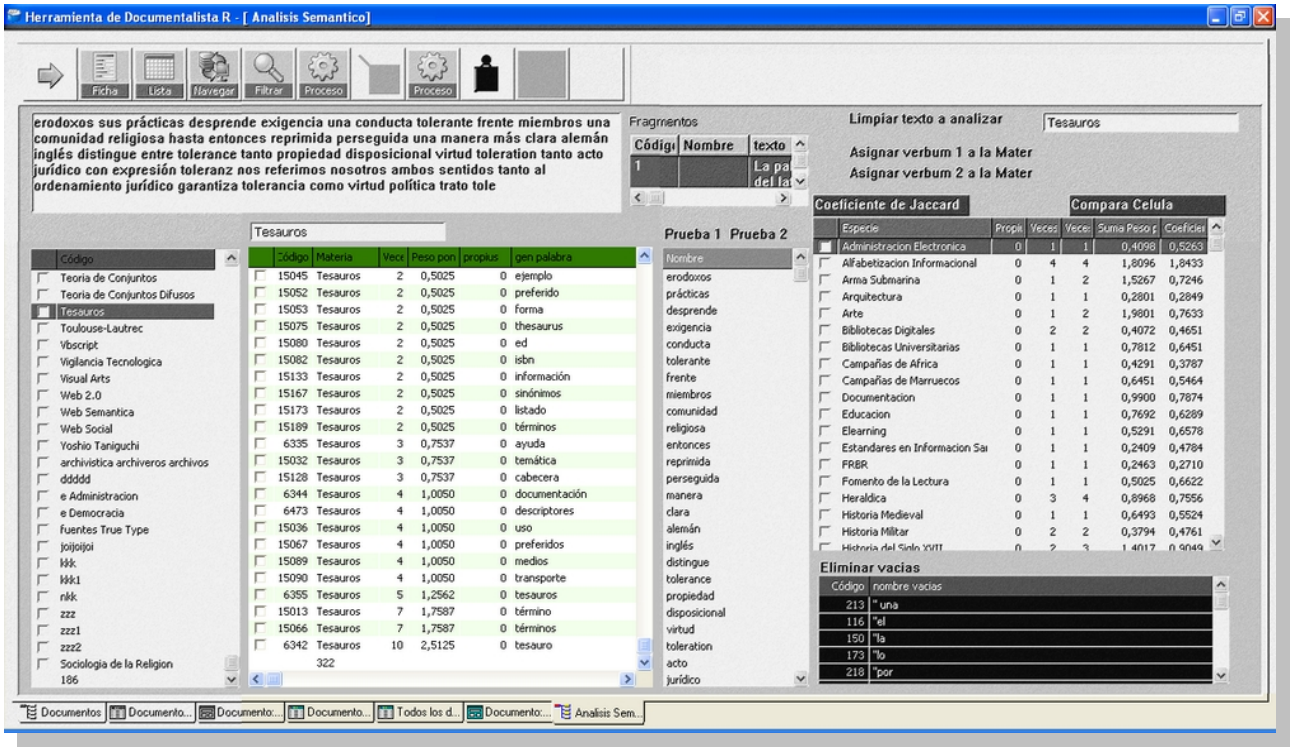
Creacion del corpus

Se establecen una serie de materias a las que se van asociando palabras (a partir de ahora utilizaremos el termino **verbum** para significar estas palabras).

Se seleccionan un grupo de documentos, se procura que estos sean lo mas significativo posibles y se se extraen los verbum de cada documento. Estos verbum se asocian a una materia. Cada materia va a tener asociada una serie de verbum. Asi por ejemplo obtenemos la siguiente lista de un documento cuya materia asignada es la “logica difusa”:

Verbum	Veces	P. Pond	Verbum	V	VP	Verbum	V	VP	Verbum	V	VP
quiere	1	0.9523	índice	1	0.9523	regla composicional	1	0.9523	tema	3	2,8571
describir	1	0.9523	objetivos:	1	0.9523	inferencia índice	1	0.9523	razonamiento	3	2,8571
humano	1	0.9523	repasar	1	0.9523	tema	1	0.9523	aproximado	3	2,8571
fuzzy	1	0.9523	reglas	1	0.9523	lógica difusa	1	0.9523	inferencia	3	2,8571
logic	1	0.9523	comprender	1	0.9523	razonamiento aproximado	1	0.9523	lógica	4	3,8095
utiliza	1	0.9523	generalización	1	0.9523	objetivos: repasar	1	0.9523	binarios	4	3,895
módulo	1	0.9523	proposiciones	1	0.9523	reglas inferencia	1	0.9523	lógica	5	4,7619
ii:	1	0.9523	difusas	1	0.9523	básicas comprender	1	0.9523	difusa	5	4,7619
fundamentos	1	0.9523	módulo ii:	1	0.9523	su generalización	1	0.9523			
razonamiento	1	0.9523	fundamentos lógica	1	0.9523	proposiciones difusas	1	0.9523			
clásica	1	0.9523	difusa tema	1	0.9523	tradicional	2	19.047			
reglas	1	0.9523	aproximado tema	1	0.9523	basada	2	1,9047			
principios	1	0.9523	aproximado 1	1	0.9523	valores	2	1,9047			
básicos	1	0.9523	razonamiento lógica	1	0.9523	VERDADERO	2	1,9047			
difusa	1	0.9523	clásica reglas	1	0.9523	y falso	2	1,9047			
modus	1	0.9523	inferencia básicas	1	0.9523	a veces	2	1,9047			
ponens	1	0.9523	principios básicos	1	0.9523	inadecuada	2	1,9047			
generalizado	1	0.9523	razonamiento lógica	1	0.9523	razonamiento	2	1,9047			
regla	1	0.9523	difusa 1	1	0.9523	básicas	2	1,9047			
composicional	1	0.9523	modus ponens	1	0.9523	lógica	2	1,9047			
índice	1	0.9523	generalizado 2	1	0.9523	difusa razonamiento	2	1,9047			

Asi tenemos que cada materia tiene asociada una serie de verbum, el numero de veces que se repite en un texto y el numero de veces ponderado al 100%.



Con este criterio, si el verbum "**razonamiento"** aparece 3 veces, tiene un aparicion ponderada de 2,8571. Este valor va a permitir que idependientemente del numero de verbum recogidos podamos comparar com mayor rigor y podamos saber cual es la importancia relativa de un verbum dentro de esa materia. De esta manera se va a paliar, entre otros problemas, el de tener que recoger el mismo numero de verbum por cada materia.

Calculo del peso ponderado
S: Sumatorio de veces totales **V**: Veces que aparece ese verbum en esa materia ***peso ponderado*** = (**V** / **S**) * 100

Código	Materia	Vece	Peso pon	propius	gen palabra
21933	ISO 23081	13	2,5048		gestión
21952	ISO 23081	13	2,5048		normas
22702	Web Semantica	13	2,8138	Propius	conocimiento
32124	Criptografia	14	2,5735	Propius	criptografía
18810	Gramatica	14	4,0229	Propius	gramática
21954	ISO 23081	14	2,6974		une
20566	Metadatos	14	3,5087	Propius	metadatos
30898	Mineria Web	14	1,9607	Propius	mineria
15416	Servicios Tributarios	14	6,5116	0	valor
21702	Vigilancia Tecnologica	14	2,6217	Propius	prospectiva
21706	Vigilancia Tecnologica	14	2,6217	Propius	tecnológica
16011	Fomento de la Lectura	15	7,5376	Propius	fomento
13960	Igualdad de Oportunidades	15	3,2608	0	vida
20956	Libros Electronicos	15	2,4350		adobe
20787	Libros Electronicos	15	2,4350	Propius	ebook
20573	Metadatos	15	3,7593	Propius	documentos
3295	Bibliotecas	16	2,4132	Propius	biblioteca
31423	Marketing Archivistico	16	3,5087	Propius	información
31411	Marketing Archivistico	16	3,5087	Propius	servicios
22395	Ontologias	16	3,1496		oac
	22093				22093

Previo a la extraccion de los verbum, se realizan las siguientes operaciones:

1) Pasar a minusculas todo el texto.

2) Eliminar signos y simbolos como comillas, signos de interrogacion, puntos etc...

3) Elimar palabras vacias.
   En una tabla vamos añadiendo las palabras vacias, que no se tendran en cuenta a la hora de realizar el corpus, cuestion no tan trivial como pudiera parecer a primera vista:

Un problema, aun sin resolver, radica en que hay palabras vacias para una materia y sin embargo son muy significativas para otras:

Una fecha (1 de agostos, 3 de diciembre) en un texto de fisica nuclear deberia ser considerada vacia, podria ser la fecha de publicacion u otro valor que nada tiene que ver con el contenido semantico del documento. Sin embargo "mayo" o mejor aun "2 de mayo" es altamente significativa si el texto trata sobre la Guerra de Independecia.

Actualmente en la version que trabajamos no contemplanos palabras vacias por materia, sino que las palabras vacias se aplican a todo el corpus. Entre otras cosas porque si desconocemos la materia del documento analizado, que es lo que queremos calcular ¿como podemos saber que palabras son vacias para ese documento? Asi en el programa piloto tenemos entre otras, palabras vacias como estas:

ya	del	algo	son	ello	a la
	por	así	dos	y del	de las
lo	se	es	muy	otro	por su
el	han	quieren	pero	unos	allá
	al	o	sea	iba	lo que
	que	estar	cual	etc	que un
la	este	solo	otras	ellas	todos
si	.	co	sus	ni	y la
una	,	para	tan	aquélla	con que
en	gran	sido	antes	aquella	cuanto
un	actualidad	ha	fuesen	otros	fu
de	etc.,	más	esa	otros que	dar
y	e	una	todo	va	vayan
no	em as,	está	n°	l a	fo
las	como	entre	dentro	los	es una
los	hacer	siguientes	después	luego	el"
a	hasta	aquellos	desde	les	qué
con	la	lleva	ser	uno	nos
..xxx	su	cuyo	a partir	donde	dice
con	sin	era	ahora	ido	que sean

La eliminacion de palabras vacias tiene dos ventajas, la principal consiste en evitar en el calculo las coincidencias con palabras que no tienen ningun significado y otra reducir el numero de palabras asociadas a cada materia para disminuir el tiempo de proceso.

Un ejemplo:

Esta operación es valida tanto para calcular la materia como para añadir al corpus materia-verbum los resultados obtenido.

Elegimos un fragmento del documento “Pautas para Bibliotecas Públicas: Preparadas por la Sección de Bibliotecas Públicas de la Federación Internacional de Asociaciones de Bibliotecas y Bibliotecarios (FIAB) Biblioteca Municipal de Peñaranda de Bracamonte Fundación Germán Sánchez Ruipérez 1998”

Texto a procesar
La Sección de Bibliotecas Públicas de la FIAB publicó finalmente en 1973 las Normas para Bibliotecas Públicas, que se reimprimieron con pequeñas correcciones en 1977.

Desde entonces se han producido muchos cambios, en todo el mundo, tanto en cuanto a recursos disponibles para eldesarrollo de las bibliotecas públicas como en cuanto a esperanzas públicas en los servicios bibliotecarios. Por ello, la Sección consideró que era el momento de examinar nuevamente estas «normas» y nombró un grupo de trabajo para este fin en 1983. Inevitablemente, el trabajo del grupo se ha realizado por medio de correspondencia principalmente y cabe suponer que ninguno de los miembros esté de acuerdo con todos y cada uno de lo s puntos de las Pautas resultantes . El in forme del grupo de trabajo se presentó a la Conferencia General de la FIAB en 1985.

Se procesa para eliminar signos, retornos de carro....

Texto sin signos

```
la sección bibliotecas públicas fiab publicó finalmente 1973 normas para bibliotecas públicas
reimprimieron con pequeñas correcciones 1977 desde entonces han producido muchos cambios todo mundo
tanto cuanto recursos disponibles para eldesarrollo bibliotecas públicas como cuanto esperanzas
públicas servicios bibliotecarios por ello sección consideró era momento examinar nuevamente estas
«normas» nombró un grupo trabajo
```

Se extraen los verbum:

1973	estas	públicas
1977	examinar	públicas
bibliotecarios	fiab	públicas
bibliotecas	finalmente	publicó
bibliotecas	grupo	recursos
bibliotecas	han	reimprimieron
cambios	la	sección
como	momento	sección
con	muchos	servicios
consideró	mundo	tanto
correcciones	nombró	todo
cuanto	normas	trabajo
cuanto	normas	un
desde	nuevamente	
disponibles	para	
eldesarrollo	para	
ello	pequeñas	
entonces	por	
era	producido	
esperanzas	públicas	

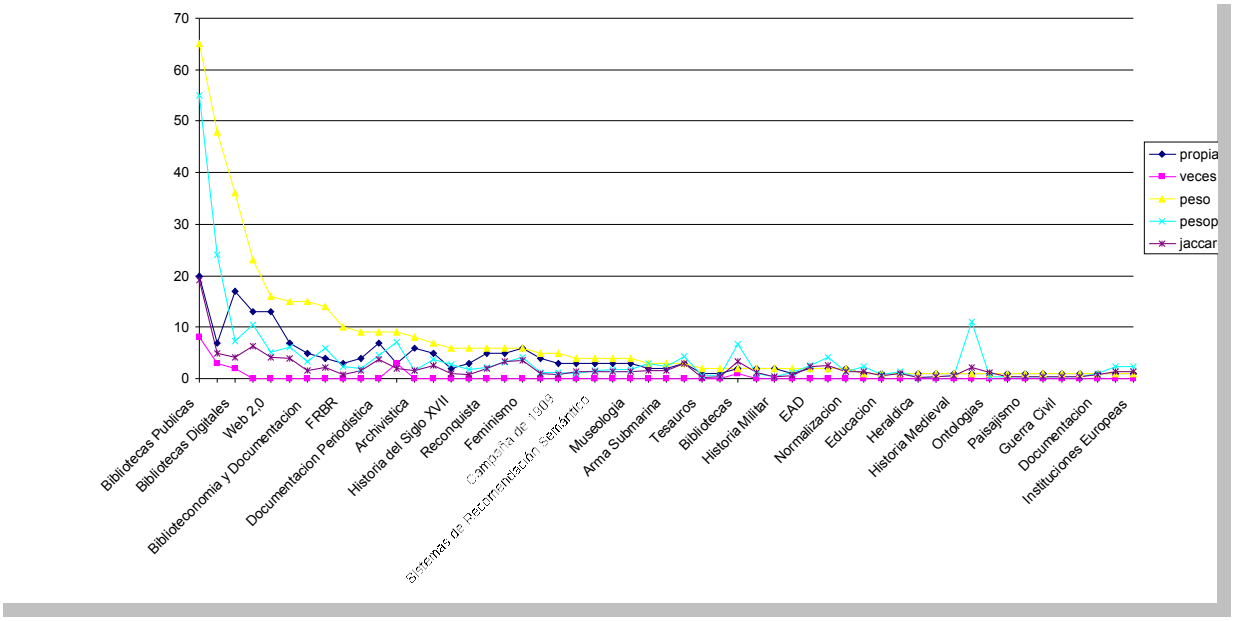
Se eliminan las vacias:

1973	nombró
1977	normas
bibliotecarios	normas
bibliotecas	nuevamente
bibliotecas	pequeñas
bibliotecas	producido
cambios	públicas
consideró	públicas
correcciones	públicas
disponibles	públicas
eldesarrollo	publicó
entonces	recursos
esperanzas	reimprimieron
examinar	sección
fiab	sección
finalmente	servicios
grupo	trabajo
momento	
muchos	
mundo	

Calculamos la semejanza contra el corpus de verbum previamente establecido, obteniendo el siguiente resultado.

Materia	veces	propius	peso	pesopond	jaccard
Bibliotecas Publicas	20	8	65	55,0834	19,0476
Fomento de la Lectura	7	3	48	24,1203	4,8951
Bibliotecas Digitales	17	2	36	7,3306	4,1162
Alfabetizacion Informacional	13	0	23	10,4062	6,3106
Web 2,0	13	0	16	5,0306	4,1401
Administracion Electronica	7	0	15	6,1473	3,8461
Biblioteconomia y Documentacion	5	0	15	3,448	1,5923
OAI	4	0	14	5,8576	2,1857
FRBR	3	0	10	2,463	0,8219
Igualdad de Oportunidades	4	0	9	1,9562	1,6393
Documentacion Periodistica	7	0	9	4,5918	3,6649
Bibliotecas Universitarias	3	3	9	7,0311	1,9867
Archivistica	6	0	8	1,6528	1,5584
Servicios de Referencia	5	0	7	3,804	2,5906
Historia del Siglo XVII	2	0	6	2,8036	0,9132
Arquitectura	3	0	6	1,6806	0,8645
Reconquista	5	0	6	2,1504	1,9011
Elearning	5	0	6	3,1746	3,4246
Feminismo	6	0	6	4,1094	3,5714
La Inquisicion	4	0	5	1,0866	0,909
Campaña de 1909	3	0	5	1,0892	0,7653
Estandares en Informacion Sanitaria	3	0	4	0,9637	1,4634
Sistemas de Recomendación Semántico	3	0	4	1,5685	1,3698

Innovacion	3	0	4	1,8432	1,3574
Museologia	3	0	4	1,6805	1,282
Arte	2	0	3	2,9701	1,5625
Arma Submarina	2	0	3	2,29	1,4814
Mujeres e Informacion	3	0	3	4,3476	3,0303
Tesauros	1	0	2	0,4987	0,2785
Imperio Romano	1	0	2	0,6269	0,3521
Bibliotecas	2	1	2	6,6666	3,3898
Servicios Tributarios	2	0	2	0,9302	1,2422
Historia Militar	2	0	2	0,3794	0,4784
Sociedad de la Informacion	1	0	2	1,1834	0,6172
EAD	2	0	2	2,6314	2,2988
Descripcion Archivistica	2	0	2	4,0816	2,4691
Normalizacion	2	0	2	1,3332	1,3513
Religion	1	0	1	2,2727	1,4084
Educacion	1	0	1	0,7692	0,6369
Geografia	1	0	1	1,4705	0,9803
Heraldica	1	0	1	0,2242	0,2518
Historia Clinica	1	0	1	0,4149	0,4219
Historia Medieval	1	0	1	0,6493	0,5586
Preservacion	1	0	1	11,1111	2,2222
Ontologias	1	0	1	0,4237	1,0869
Lexicografia	1	0	1	0,3344	0,3344
Paisajismo	1	0	1	0,4347	0,4464
Economia	1	0	1	0,4566	0,4672
Guerra Civil	1	0	1	0,3937	0,4
Teoria de Conjuntos Difusos	1	0	1	0,3184	0,3546
Documentacion	1	0	1	0,99	0,8
Seguridad	1	0	1	2,2727	1,3698
Instituciones Europeas	1	0	1	2,2727	1,3698



En este caso la materia probable es evidentemente “Bibliotecas Publicas”, en el siguiente grafico la eleccion es intuitiva.

Los valores mas altos corresponden a la materia señalada. A continuacion definiremos con mas detalle los criterios utilizados en el calculo de los valores que aparecen en los resultados:

Materia	C1 propius	C2 veces	C3 peso	C4 Peso pond.	C5 Jaccard
Bibliotecas Publicas	8	20	65	55,0834	19,0476
Fomento de la Lectura	3	7	48	24,1203	4,8951
Bibliotecas Digitales	2	17	36	7,3306	4,1162

Definicion de los criterios utilizados

C1 propius

Este valor se añadio recientemente, en la version HDD 3.0 86 para diferenciar los verbum que aparecen muchas veces, frente los verbum muy significativos. Por regla general suelen aparecer con mas frecuencias los terminos significativos, pero no siempre es asi. Es mas, terminos muy relevantes y significativos pueden aparecer muy pocas veces o solo una vez. Como estos terminos son determinantes para discernir la materia a la que pertenece un documento, se permite señalar en el corpus los verbum, que hemos llamado **“propius”**.

Ejemplo: Extraidos los verbum de un documento sobre “referencias bibliográficas” el verbum “vancouver” solo aparecia una vez, como consideramos que Vancouver es el nombre de un estilo de referencias bibliográficas; Señalariamos “vancouver” como verbum propius.

C2 veces

Es el numero de veces que aparece un verbum de la muestra en el corpus y en esa materia, en nuestro ejemplo de los verbum de la muestra aparecen 20 veces en la materia “Bibliotecas Publicas” , 17 veces en “Bibliotecas Digitales” y 7 veces en la materia “Fomento de la lectura”

C3 peso

Esta relacionado con veces, el peso se calcula a partir de las ocurrencias de un verbum y el numero de veces que aparece ese verbum en el corpus.

C4 pesopond

Es el peso ponderado, en tanto por ciento respecto a la materia.

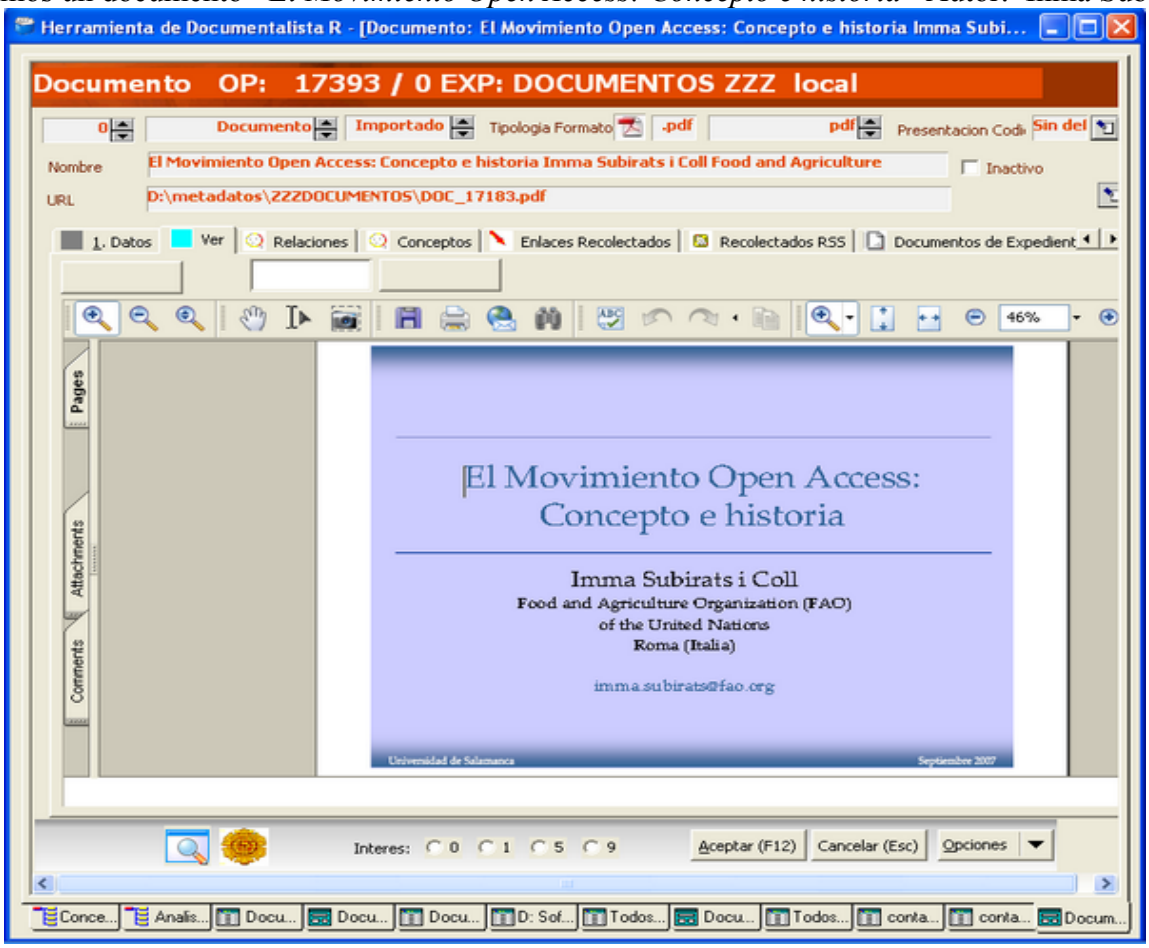
C5 Jaccard

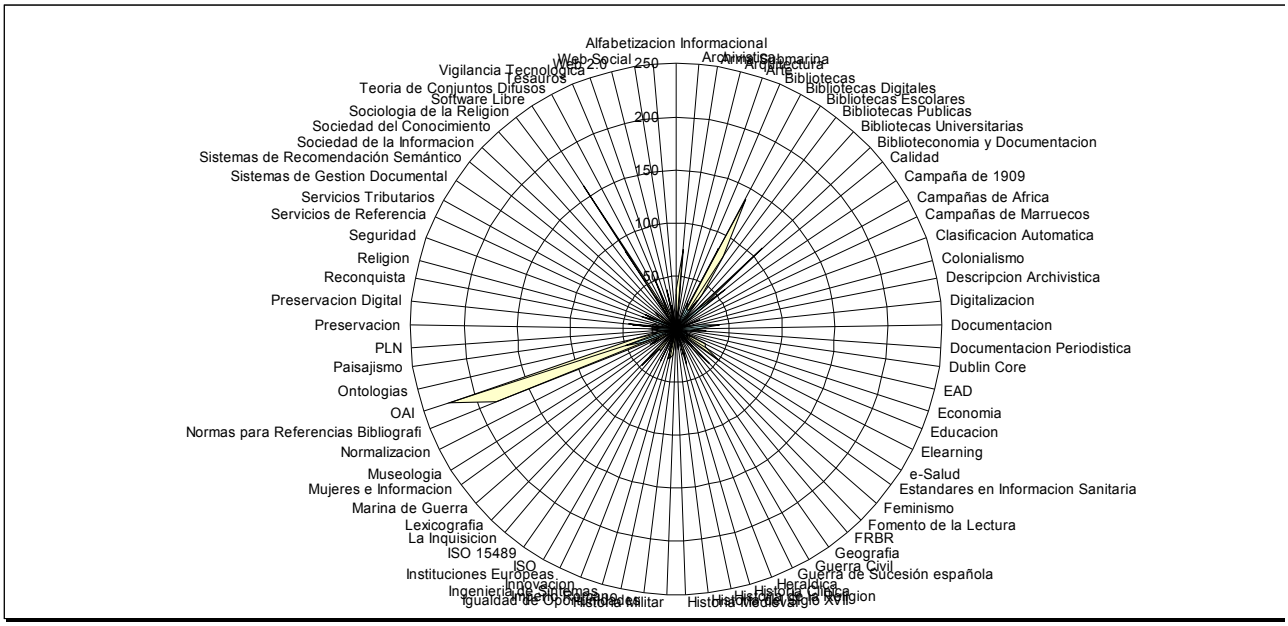
Basado en la medida de la similitud de Jaccard, el calculo de la similitud entre dos conjuntos X e Y se deduce por el cociente entre el numero de elementos de la interseccion y el numero de elementos de la union.

¿Que materia elegir cuando los resultados no son tan evidentes?

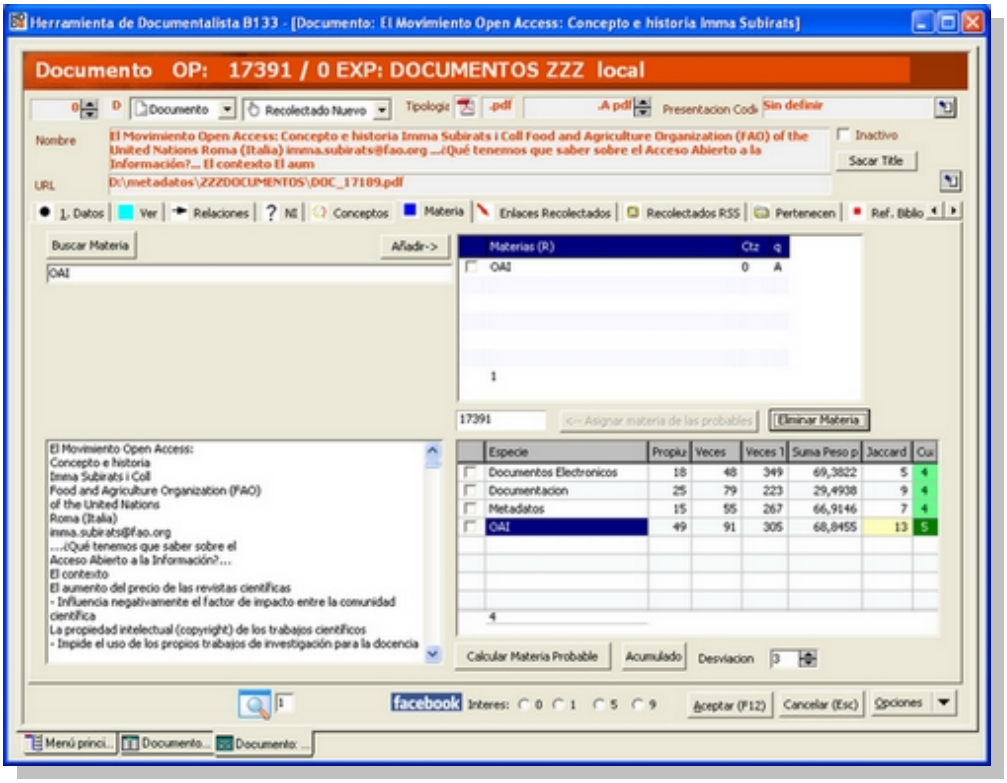
No siempre los resultados son como los del ejemplo anterior, De echo puede darse un alto coeficiente de Jaccard pero un bajo peso ponderado o cualquier otra posibilidad. A traves de un numero de pruebas se ha observado que la materia probable suele coincidir con los primeros valores de cada uno de los criterios.

Elegimos un documento *"El Movimiento Open Access: Concepto e historia"* Autor: Inma Subirats





Ante la dificultad de elegir la materia probable, hemos añadido un sexto criterio que permite señalar las materias con valores mas altos en cada uno de los criterios. asi tenemos, para el documento *"El Movimiento Open Access: Concepto e historia"* que la materia *OAI* (*Open Access Initiative)* aparece 5 veces entre los criterios con mas valor.



Materia	c1	c2	c3	c4	c5	Vi
OAI	71	47	224	50,5621	9,6205	5
Bibliotecas Digitales	86	22	139	20,9334	8,3252	4
Normas para Referencias Bibliografi	64	19	183	35,2565	7,9601	4
Biblioteconomia y Documentacion	56	0	111	24,0759	6,5497	3
Software Libre	88	2	160	36,3586	10,6024	3
Alfabetizacion Informacional	32	9	37	16,8942	4,2609	2
Calidad	10	8	26	59,0902	1,6051	2
Digitalizacion	17	3	19	41,3041	2,707	2
e-Salud	28	21	32	11,3472	3,7735	2
Administracion Electronica	25	0	63	25,8183	3,4246	1
Archivistica	40	4	75	15,495	4,362	1
Bibliotecas Escolares	53	0	80	21,68	6,3855	1
Bibliotecas Publicas	16	4	40	33,8974	2,3703	1
Documentacion	19	0	19	18,81	2,8231	1
Documentacion Periodistica	21	0	28	14,2856	2,8263	1

Dublin Core	17	0	17	16,5036	2,5147	1
Elearning	22	0	30	15,873	3,1654	1
Estandares en Informacion Sanitaria	19	0	51	12,2882	2,5165	1
Fomento de la Lectura	19	2	45	22,6127	2,7259	1
Historia Clinica	25	0	34	14,1066	3,2092	1
Historia del Siglo XVII	18	0	20	9,3442	2,3407	1

El calculo se hace ordenando la tabla de resultados por cada uno de los criterios y asignando un valor de verdadero o falso (1,0) si aparece entre los valores mas altos, los que son mayores que la media de ese criterio. El nuevo criterio toma una puntuacion de 0 a 5 donde 0 (cero) indica que nunca ha aparecido entre los valores mas altos de ningun criterio, 1 (uno) que aparecio una vez entre los valores mas altos y asi sucesivamente, para el documento modelo la materia *OAI (Open Access Initiative)* siempre aparece en los primeros puestos de las diversos criterios.

OAI	71	47	224	50,5621	9,6205	**5**
Bibliotecas Digitales	86	22	139	20,9334	8,3252	**4**
Normas para Referencias Bibliografi	64	19	183	35,2565	7,9601	**4**
Biblioteconomia y Documentacion	56	0	111	24,0759	6,5497	**3**
Software Libre	88	2	160	36,3586	10,6024	**3**
Alfabetizacion Informacional	32	9	37	16,8942	4,2609	**2**
Calidad	10	8	26	59,0902	1,6051	**2**
Digitalizacion	17	3	19	41,3041	2,707	**2**
e-Salud	28	21	32	11,3472	3,7735	**2**

Ejemplo de calculo del Criterio sexto CVI.

A continuacion se puede ver un ejemplo simplificado basado en un documento ficticio con unos verbum elegidos al azar, donde se observa con mas claridad la forma de realizar el calculo de este criterio:

Una vez que se han realizado todos los procesos de calculo de criterios descritos anteriormente, se ordena la tabla de resultados por el criterio "1 propius", se calcula el valor medio y se seleccionan los valores que son mayores que el valor medio para ese criterio. En este caso corresponde a las materias *"Bibliotecas Universitarias", "Paisajismo"* y "*Preservacion Digital*". La media de la columna es 1,3. Asi se seleccionan los que tienen un valor igual a 2. (2>0,3 y 1<0,3)

Materia	C1 Propius	C2 Veces	C3 Veces Total	C4 Suma Ponderada	C5 Jaccard
Bibliotecas Universitarias	**2**	2	3	2,3437	1,6949
Paisajismo	**2**	1	5	1,739	1,0582
Preservacion Digital	**2**	0	2	0,6134	0,7017
Calidad	1	0	1	2,2727	3,125
Descripcion Archivistica	1	0	1	2,0408	2,0833

Vigilancia Tecnologica	1	0	3	2,5423	1,1904
Arte	1	0	2	1,9801	1,0526
Educacion	1	0	1	0,7692	0,813
Estandares en Informacion Sanitaria	1	0	1	0,2409	0,578
Tesauros	1	0	1	0,2512	0,3086
Media	1,3	0,3	2	1,4	1,2

Se reordena por el criterio “2 veces” donde *“Bibliotecas Universitarias” y “Paisajismo”* tienen un valor mayor que la media. La media de este columna es 0,3 donde 2>0,3 y 1>0,3

Materia	C1 Propius	C2 Veces	C3 Veces Total	C4 Suma Ponderada	C5 Jaccard
Bibliotecas Universitarias	2	**2**	3	2,3437	1,6949
Paisajismo	2	1	5	1,739	1,0582
Preservacion Digital	2	0	2	0,6134	0,7017
Calidad	1	0	1	2,2727	3,125
Descripcion Archivistica	1	0	1	2,0408	2,0833
Vigilancia Tecnologica	1	0	3	2,5423	1,1904
Arte	1	0	2	1,9801	1,0526
Educacion	1	0	1	0,7692	0,813
Estandares en Informacion Sanitaria	1	0	1	0,2409	0,578
Tesauros	1	0	1	0,2512	0,3086
Media	1,3	0,3	2	1,4	1,2

Se repite el calculo y se ordena por el criterio “C3 veces total”

Materia	C1 Propius	C2 Veces	C3 Veces Total	C4 Suma Ponderada	C5 Jaccard
Paisajismo	2	1	**5**	1,739	1,0582
Bibliotecas Universitarias	2	2	**3**	2,3437	1,6949
Vigilancia Tecnologica	1	0	**3**	2,5423	1,1904
Preservacion Digital	2	0	2	0,6134	0,7017
Arte	1	0	2	1,9801	1,0526
Calidad	1	0	1	2,2727	3,125
Descripcion Archivistica	1	0	1	2,0408	2,0833
Educacion	1	0	1	0,7692	0,813
Estandares en Informacion Sanitaria	1	0	1	0,2409	0,578
Tesauros	1	0	1	0,2512	0,3086
Media	1,3	0,3	2	1,4	1,2

Se ordena por el criterio “C4 suma ponderada” y se repite el calculo

Materia	C1 Propius	C2 Veces	C3 Veces Total	C4 Suma Ponderada	C5 Jaccard
Vigilancia Tecnologica	1	0	3	**2,5423**	1,1904
Bibliotecas Universitarias	2	2	3	**2,3437**	1,6949
Calidad	1	0	1	**2,2727**	3,125
Descripcion Archivistica	1	0	1	**2,0408**	2,0833
Arte	1	0	2	**1,9801**	1,0526
Paisajismo	2	1	5	**1,739**	1,0582
Educacion	1	0	1	0,7692	0,813

Preservacion Digital	2	0	2	0,6134	0,7017
Tesauros	1	0	1	0,2512	0,3086
Estandares en Informacion Sanitaria	1	0	1	0,2409	0,578
Media	1,3	0,3	2	1,4	1,2

Por ultimo se ordena por el criterio “C5 Coeficiente de Jaccard” y se realiza la misma operación de calculo.

Materia	C1 Propius	C2 Veces	C3 Veces Total	C4 Suma Ponderada	C5 Jaccard
Calidad	1	0	1	2,2727	**3,125**
Descripcion Archivistica	1	0	1	2,0408	**2,0833**
Bibliotecas Universitarias	2	2	3	2,3437	**1,6949**
Vigilancia Tecnologica	1	0	3	2,5423	1,1904
Paisajismo	2	1	5	1,739	1,0582
Arte	1	0	2	1,9801	1,0526
Educacion	1	0	1	0,7692	0,813
Preservacion Digital	2	0	2	0,6134	0,7017
Estandares en Informacion Sanitaria	1	0	1	0,2409	0,578
Tesauros	1	0	1	0,2512	0,3086
Media	1,3	0,3	2	1,4	1,2

El resultado final es que la materia *“Bibliotecas Universitarias”* tiene un valor de 5 para el criterio VI, lo que indica que estaba entre las primeras posiciones por encima de la media de todos los criterios. Esta por tanto, seria la materia probable elegida para este documento ficticio.

Resultado Final						
Materia	C1 Propius	C2 Veces	C3 Veces Total	C4 Suma Ponderada	C5 Jaccard	**VI**
Bibliotecas Universitarias	2	2	3	2,3437	1,6949	**5**
Paisajismo	2	1	5	1,739	1,0582	4
Calidad	1	0	1	2,2727	3,125	2
Descripcion Archivistica	1	0	1	2,0408	2,0833	2
Vigilancia Tecnologica	1	0	3	2,5423	1,1904	2
Arte	1	0	2	1,9801	1,0526	1
Preservacion Digital	2	0	2	0,6134	0,7017	1
Educacion	1	0	1	0,7692	0,813	0
Estandares en Informacion Sanitaria	1	0	1	0,2409	0,578	0
Tesauros	1	0	1	0,2512	0,3086	0

Tamaño de la muestra y certeza de la materia probable

En principio cuanto mayor sea la muestra (verbum extraidos del documento que se quiere analizar) la probabilidad de que la materia elegida como probable sea la correcta aumenta.

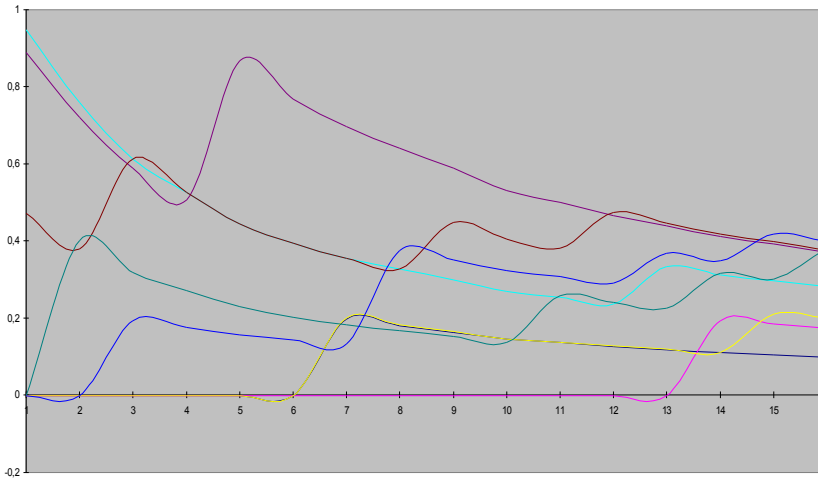
Pero como por una parte aumentar la muestra supone mayor tiempo de proceso y por otra un documento tiene un numero finito de verbum extraibles. Es necesario saber si los resultados son o no confiables.

Se ha analizado el comportamiento del criterio “Coeficiente de Jaccard” y se ha observado la evolucion y tendencia de este valor respecto al tamaño de la muestra elegida.

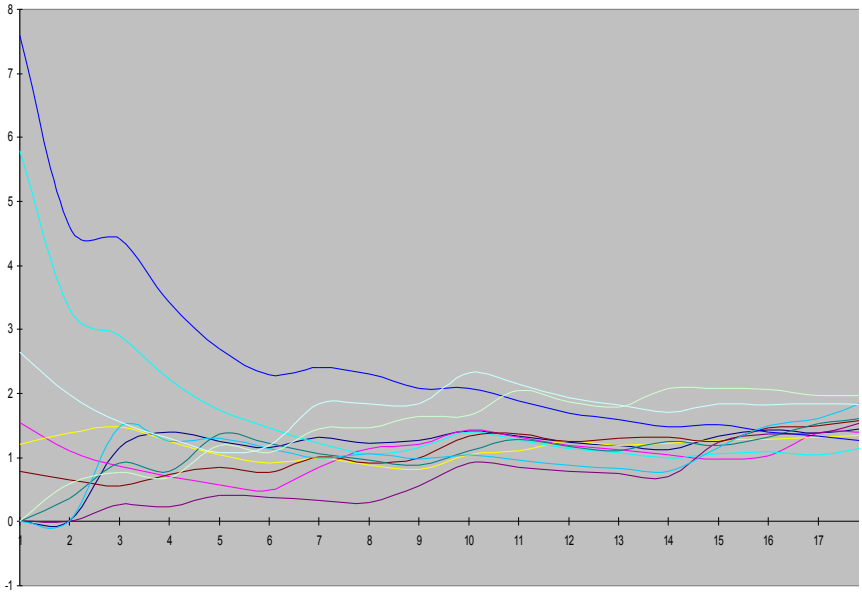
Se ha observado que puede establecerse un limite en el tamaño de la muestra ya que la semejanza de una muestra, extraida a partir de un documento , respecto al corpus preestablecido de una materia no aumenta ni disminuye de una manera relevante. O al menos los resultados que se obtienen son similares.

A continuacion puede observarse en los tres graficos siguientes como evoluciona el valor del coeficiente de Jaccard para una materia respecto a un mismo documento.

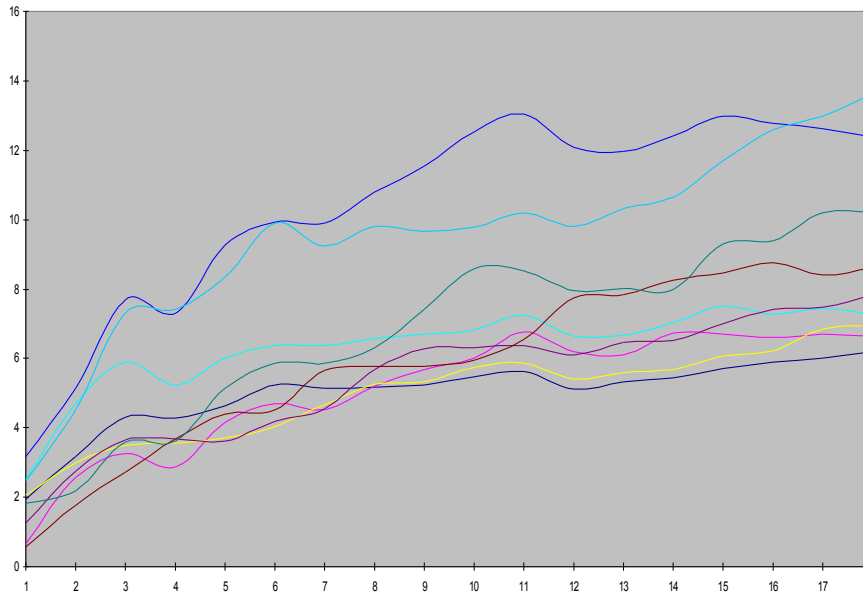
Evolucion del coeficiente de Jaccard para las materias con la probabilidad mas baja de ser las correctas. Observese la clara tendencia a la baja:



Evolucion del coeficiente de jaccard para las materias con la probabilidad media de ser las correctas, observese como se van concentrando los valores en un area.



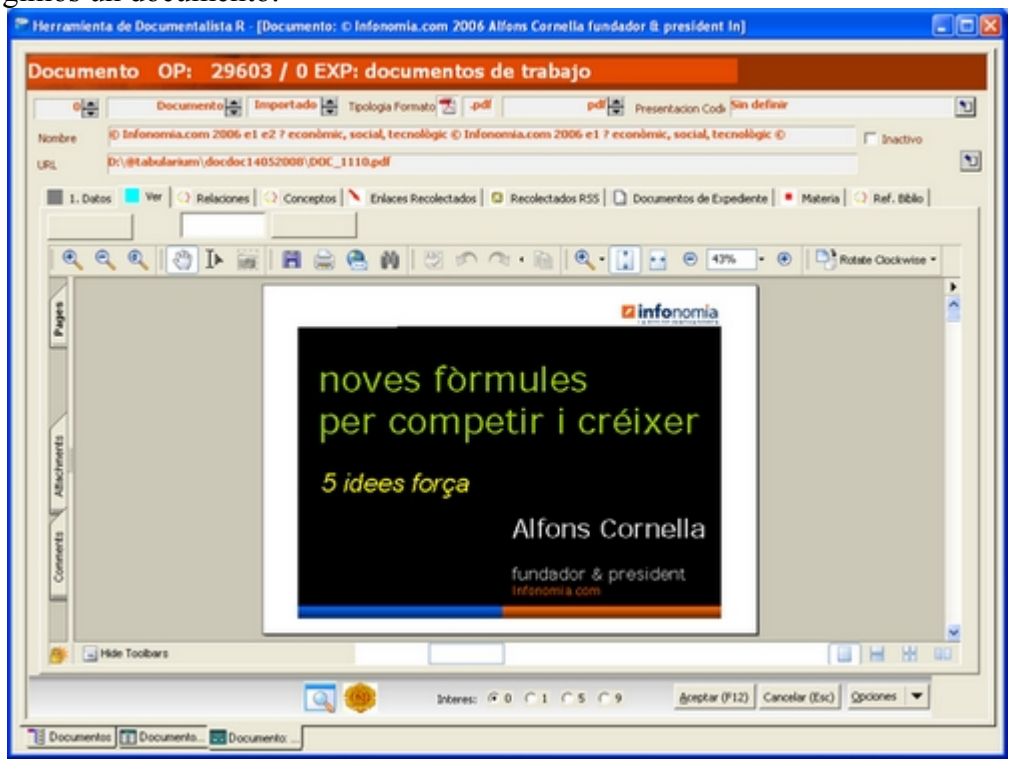
Evolucion del coeficiente de Jaccard para las materias con la probabilidad mas alta de ser las correctas, la tendencia es que los valores vayan aumentando.



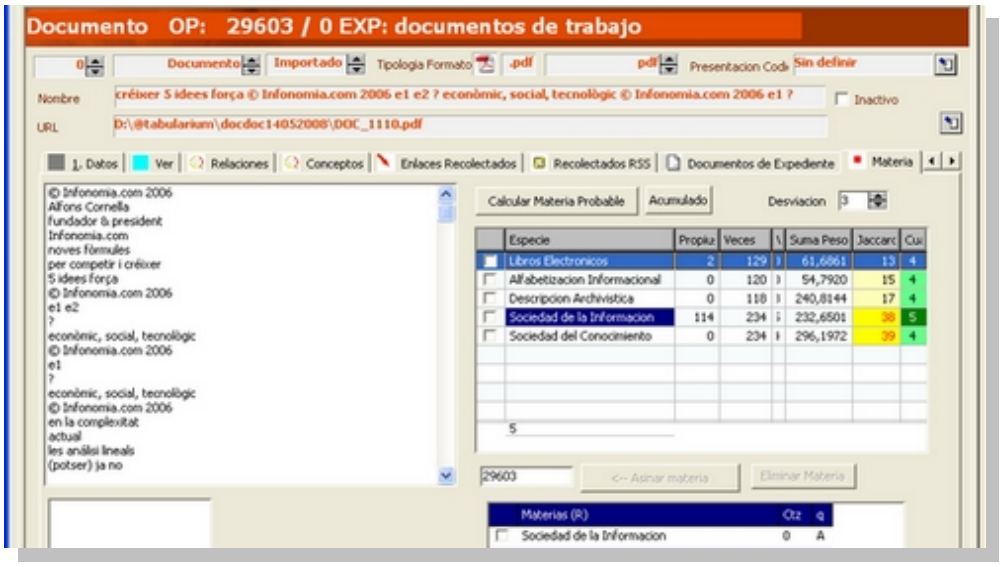
Esto va a permitir optimizar el calculo de la materia, aumentar la seguridad en que la opcion elegida ha sido la adecuada y en algunos casos evitar asignar una materia que no corresponde al documento seleccionado.

Un ejemplo de calculo de la materia probable, automatizado con HDD.

Elegimos un documento:



Realizamos el calculo en el formulario correspondiente: La materia se asigna de forma automatica si se elige esa opcion, o bien se selecciona manualmente entre las opciones que nos muestra el programa.



Valores totales obtenidos respecto al documento anterior, el mayor valor de VI es 5, por lo tanto la materia probable seleccionada seria “Sociedad de la Informacion”.

	C1	C2	C3	C4	C5	VI
Libros Electronicos	129	2	380	616.861	13	4
Sociedad de la Informacion	234	114	235	2.326.501	38	5
Sociedad del Conocimiento	234	0	234	2.961.972	39	4
Servicios Tributarios	12	0	155	720.927	1	2
Metadatos	120	1	131	328.288	14	3
Alfabetizacion Informacional	120	0	120	547.920	15	4
Descripcion Archivistica	118	0	118	2.408.144	17	4
Arquitectura de la Informacion	36	0	69	181.565	3	1
HL7	12	0	52	0	1	0
Web Semantica	46	0	47	101.709	4	1
Representacion y Organizacion del C	22	0	44	118.908	2	1
Preservacion Digital	20	1	43	131.892	1	0
Documentacion	12	0	43	56.870	1	0
Teoria de Conjuntos Difusos	13	0	35	116.270	1	0
Igualdad de Oportunidades	16	0	32	69.552	1	0
Biologia del Cerebro	19	0	32	88.141	2	1
Vigilancia Tecnologica	10	0	24	44.940	0	0
Bibliotecas Digitales	12	2	21	31.626	1	0
Gestion de Documentos Audiovisuales	18	0	20	49.260	1	0
Clasificacion Automatica	7	3	19	63.542	0	0
Normalizacion	15	0	19	60.500	1	0